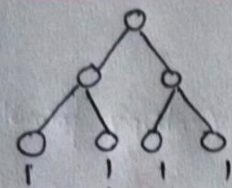
(15)

ENTROPY IN DECISION TREE



Leaf Node
(FINAL CLASS LABELS)

- Which feature to select first, we use entropy.
- Suppose we have 3 feature f_1, f_2, and f_3.
  So which feature to choose first. f_1 / f_2 / f_3
- Entropy ranges from 0 to 1. (0 - Pure split. 3yes 0no; 1 - Means basic impure subset equal portion. 3Yes, 3NO)
- Entropy only used for one node.
- Entropy = 0, then it is leaf node.

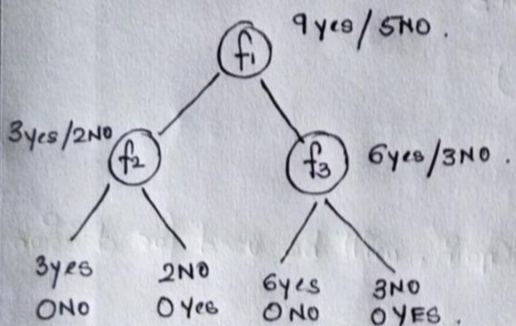
- Entropy measure the purity of splits.

$$\boxed{H(S) = -P_{(+)} \log_2 (P_+) - P_{(-)} \log_2 P_{(-)}}$$

P_+ = % of +ve class
probability of +ve class.

P_- = % of -ve class
probability of -ve class.

S = Sample of Training Examples.
(Subset)



Let find out entropy for F_2. (3yes/2NO)

$$= -\frac{3}{5} \log_2 \left(\frac{3}{5}\right) - \left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right)$$

$$= 0.78 \text{ bits}.$$

If we have complete impure set (equal yes no) that is worst split.

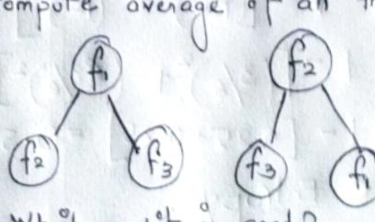
- So calculate entropy of all the variables, the variable which has lowest entropy is selected first for splitting.

But when we select a node, it is splitted into many sub-node which will also have some entropy, so we need to do summation of all entropy for that we use Information Gain.

INFORMATION GAIN

- Information gain collection all the entropy value from root node to leaf node.
- Information gain is a collection of all entropy value whereas entropy is for one node only.
- Compute average of all the entropy.

$$\boxed{Gain(S, A) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v)}$$

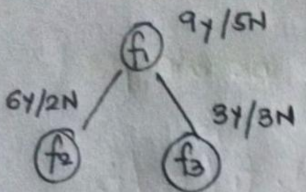


Which split is good?

$H(S_v)$ = Entropy (subset after splitting)

$H(S)$ = Entropy of all variable like entropy of f_1, entropy of f_2 and entropy of f_3

S = Subset S_v = Subset after splitting



$$H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v).$$

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_- \quad \text{(entropy formula)}$$

for f_1, $H(F_1) = H(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

(Information gain is done for root node)

$$= 0.91.$$

$$H(S_v) = H(F_2) = -\frac{6}{8}\log\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.81.$$

$$H(S_v) = H(F_3) = -\frac{3}{3}\log\left(\frac{3}{3}\right) - \frac{3}{3}\log_2\left(\frac{3}{3}\right) = 1 \quad \text{(completely impure)}$$

$$\text{Gain} = H(S) - \frac{6+2}{6+2+3+3} H(F_2) - \frac{3+3}{6+2+3+3} H(F_3)$$

$$= 0.91 - \frac{8}{14}(0.81) - \frac{6}{14}(1)$$

$$= 0.049.$$

Calculate for all combination.

Combination which is giving highest Information Gain, will be used for decision tree construction.

GINI IMPURITY IN DECISION TREE

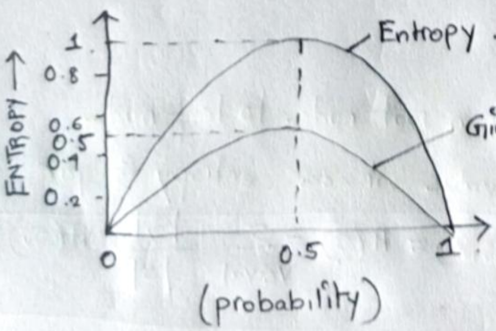
- Also calculate purity of split.

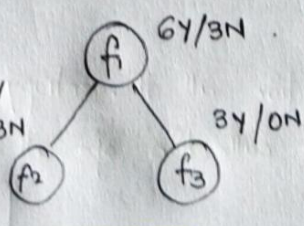
GINI INDEX

$$\boxed{GI = 1 - \sum_{i=1}^{n} (P)^2}$$

$$= 1 - \left[(P_+)^2 + (P_-)^2\right].$$

Entropy Graph →





$$GI(f_2) = 1 - \left[(P_+)^2 + (P_-)^2\right]$$

$$= 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

$$= 1 - [0.25 + 0.25] = 0.5$$

$$\text{Entropy} = -P_+ \log_2 P - P_- \log_2 P_-$$

$$E(f_2) = -\frac{3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)$$

$$= 1.$$

So, Gini Index < Entropy.

Difference between Gini Impurity and Entropy.

Gini Impurity ranges from 0 to 0.5.

Entropy range from 0 to 1.

Gini impurity are mostly used in ensemble technique like Random forest because of time complexity. Gini Index take less time than entropy.

$$\text{Entropy} = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P)^2$$

Usually, log take more time. So, entropy take more time.

End of day, use Information Gain. but Gini Impurity used before that before split.

Gini Index, Entropy and Information Gain are used only for CATEGORICAL VALUES

If ~~Target~~ Variable is Continuous Values / Numerical Values.

1) Decision tree sort all the values in Increasing Order. (Sorting all Values)
2) Set the threshold values (all values will be taken first 2.3 then 3.6 then 4 till last)
3) Disadvantage is if we increase the samples, threshold will have larger set which will increase time and its complexity.

Variable (f_1)	Output.
2.3	Yes
3.6	Yes
4	No
5.2	No
6.7	Yes
8.9	No
10.5	Yes
14.2.	Yes.

1) Increased order — Done.
2) First threshold will be 2.3.

f_1: ≤ 2.3 → 1Y/0No; > 2.3 → 3Y/4N

2.3.

Then next threshold will be 3.6

f_1: ≤ 3.6 → 2Y/0N; > 3.6 → 1Y/4N.

So. for each threshold impurity and Information gain will be calculated.

Highest Information gain will be choosed.

- Only disadvantage is if it have larger set, decision tree will take time to train. So time complexity is the issue.

7 Ginie Index Vs Information Entropy. (Decision tree)

→ Decision tree optimize each split on maximizing purity. Purity can be thought of as how homogenized the grouping are. Depending on which impurity is measured, tree classification can vary.

Entropy → If the sample is completely homogenous then entropy is zero and if the sample is equally divided it has entropy of one.

Information gain → The information gain is based on the decrease in entropy after data-set is split on attributes.

- Constructing a decision tree is all about finding attributes that returns the highest information gain. (i.e, most homogenous branches).

Gini Index → Gini index, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

1. It works with categorical variables like pass/fail.
2. It performs binary splits.
3. Higher the value of Gini higher the homogenity.
4. CART (Classification and Regression Tree) uses Gini method to create binary split.

Chi-Square → statistical significance between the difference between sub nodes and parent node.

- We measured by sum of standardised difference between observed and expected.
- Works with categorical variables.
- Can perform two or more split.
- Higher the value of Chi-Square higher the statistical significance of difference between sub node & parent node.

$$\text{chi square} = \left(\frac{(\text{Actual} - \text{Expected})^2}{(\text{Expected})}\right) / 2$$

- It generated CHAID (Chi-square automated Interaction detector)

Reduction In Variance — Used for continuous (Above 4, used for categorical)

$\text{Variance} = \frac{\Sigma(x-\bar{x})^2}{n}$, split has lower variance compared to parent node, split take place.